WHAT WORKS CLEARINGHOUSE
STUDY REVIEW STANDARDS

Revised February 2006

INTRODUCTION

The Institute of Education Sciences (IES) and the What Works Clearinghouse (WWC) have identified 10 topic areas that present a wide range of our nation's most pressing issues in education (e.g., middle school math, beginning reading, and character education). Within each selected topic area, the WWC collects studies of interventions (i.e., programs, products, practices, and policies) that are potentially relevant to the topic area through comprehensive and systematic literature searches. The studies collected are then subjected to a three-stage review process.

First, the WWC screens studies based on their relevance to the particular topic area, the quality of the outcome measures, and the adequacy of data reported. Studies that do not pass one or more of these screens are identified as "Does Not Meet Relevance Screens" and hence excluded from the WWC review.

Second, for each study that "Meets Relevance Screens," the WWC assesses the strength of the evidence that the study provides for the effectiveness of the intervention being tested. Studies that provide strong evidence for an intervention's effectiveness are characterized as "Meet Evidence Standards," studies that offer weaker evidence "Meet Evidence Standards with Reservations," and studies that provide insufficient evidence "Do Not Meet Evidence Screens." In order to meet evidence standards (either with or without reservations), a study has to be a randomized controlled trial, a regression discontinuity design, a quasi-experiment, or a single-case design.[1] The rules for determining the specific evidence category that a study falls under depends on the design of the study, as will be detailed later in the document.

At the third stage, studies that are rated as meeting evidence standards (either with or without reservations) during the second stage are reviewed further to gather information about other important characteristics, including variations in participants, settings, and outcomes; testing within subgroups, and statistical reporting. Note that the information collected from the third review stage is for descriptive purposes. It does not affect the

1 Randomized controlled trials are studies in which participants are randomly assigned to an intervention group that receives or is eligible to receive the intervention and a control group that does not receive the intervention. Regression discontinuity designs are designs in which participants are assigned to the intervention and the control conditions based on a cutoff score on a pre-intervention measure that typically assesses need or merit. This measure should be one that has a linear relationship with the outcome of interest over the range relevant for the study sample. Quasi-experimental designs are designs in which participants are not randomly assigned to groups. Single-case designs are designs that involve repeated measurement of a single subject (e.g., a student or a classroom) in different conditions or phases over time. See the WWC Study Design Classification technical working paper for details.

rating of the strength of evidence for the intervention's effectiveness determined during the second review stage.

Based on studies that "Meet Evidence Standards" and "Meet Evidence Standards with Reservations," the WWC produces two levels of reports: intervention reports and topic reports. The intervention reports summarize evidence from multiple studies on a specific intervention. Similarly, the topic reports summarize evidence for all interventions in a specific topic area.

Neither the WWC nor the U.S. Department of Education endorses any interventions.

STAGE 1: DETERMINING THE RELEVANCE OF THE STUDY TO THE WWC REVIEW

OVERVIEW

In each topic area identified by the IES and the WWC, the WWC collects both published and unpublished impact studies that are potentially relevant to the topic. The WWC review team then screens all collected studies to ensure that the studies to be included in the WWC review were conducted within an relevant timeframe, are relevant to the topic, include a sample relevant to the WWC's research question, use appropriate measures for relevant outcomes, and report findings adequately.

SCREENING STANDARDS

- **Relevant Timeframe**: The study must have been conducted during a timeframe relevant to the WWC review. For example, in the topic area of middle school math, only studies conducted after 1983 are eligible for inclusion in the WWC review.

- **Relevant Intervention**: The intervention must be relevant to the WWC review. An intervention designed to improve students' writing skills, for example, is not a relevant intervention for the topic area of beginning reading. In contrast, a study of an intervention designed to improve vocabulary would be.

- **Relevant Sample**: The study's sample must be relevant to the WWC review. In the topic area of beginning reading, for example, a relevant study sample has to consist of students in grades K-3.

- **Relevant Outcome**: The study must report on at least one outcome relevant to the WWC review. Student engagement, for example, is not considered a relevant outcome for interventions in middle school math, which focuses on achievement outcomes.

- **Adequate Outcome Measure**: The measure used must be able to reliably measure a relevant outcome that it is intended to measure.[2] For example, a nationally normed, validated test of math computation skills would be an adequate measure of math skills. In contrast, a self-report of math competency

2 The study author must provide the title of the test and one or more of the following: (1) documentation that the test items are relevant to the topic, (2) a description of the test items that is sufficient to demonstrate that the items are relevant to the topic, or (3) evidence of test reliability.

would not be considered a reliable measure of math competency.

- **Adequate Reporting**: It must be possible to calculate a standardized effect size for at least one adequate measure of a relevant outcome. In the simplest RCT design, for example, this requires the study report means of the outcomes for the intervention and comparison groups, the standard deviation of the outcome measure for the intervention and comparison groups, and the sample size for the intervention and comparison groups.
    - By default, the WWC calculates standardized effect sizes using the pooled standard deviation. If the pooled standard deviation is not available, the standard deviation for the comparison group, if available, will be used to calculate the effect sizes.
    - For studies that report effect sizes but do not provide data for computing the effect sizes, the WWC will report the effect sizes presented in the study unless there is reason to cast them in doubt.

STAGE 2: ASSESSING THE STRENGTH OF THE EVIDENCE THAT THE STUDY PROVIDES FOR THE INTERVENTION'S EFFECTIVENESS

OVERVIEW

The WWC reviews each study that passes the preceding screens to determine whether the study provides strong evidence ("Meets Evidence Standards"), weaker evidence ("Meets Evidence Standards with Reservations"), or insufficient evidence ("Does Not Meet Evidence Screens") for an intervention's effectiveness. Studies that "Meet Evidence Standards" are well-designed and implemented randomized controlled trials. Studies that "Meet Evidence Standards with Reservations" are quasi-experiments with equating[3] and no severe design or implementation problems, or randomized controlled trials with severe design or implementation problems. The evidence standards for single-case and regression discontinuity studies are under development as of January 2006.

EVIDENCE STANDARDS

Study Design: In order for a study to be rated as meeting evidence standards (with or without reservations), it must employ one of the following types of research designs: a randomized controlled trial, a quasi-experiment with equating, a regression discontinuity design, or a single-case design.

3 Equating may be done either through matching to make the study groups comparable in terms of important pre-intervention characteristics, or through statistical controls during the analysis stage to adjust for differences between the study groups, or both.

*If the study appears to be a **randomized controlled trial**, the following rules are used to determine whether the study "Meets Evidence Standards" or "Meets Evidence Standards with Reservations."*

- **Randomization**: Studies in which participants (e.g., students, teachers/classrooms, or schools) were randomly assigned to different groups are assumed to *Meet Evidence Standards*, unless one or more of the following conditions is violated:[4]

- **Baseline Equivalence**: For an RCT to *Meet Evidence Standards*, the groups should have been comparable at baseline, or incomparability should have been addressed in the analysis.
    - If there is incomparability that is not corrected for in the impact estimates reported, the study *Meets Evidence Standards with Reservations*.[5]

- **Overall Attrition**: Attrition is defined as a failure to measure the outcome variable on all the participants initially engaged in the intervention and comparison groups. High overall attrition generally makes the results of a study suspect, although there are rare exceptions. In some schools or districts, for instance, high mobility and therefore high attrition are the norm. In those cases, it could be argued that attrition problems within the study are unrelated to the intervention or to the study's internal validity, and should not cause the study to be downgraded. For an RCT to *Meet Evidence Standards,* there should not have been a severe overall attrition problem, although exceptions may exist.
    - If there was severe overall attrition that cannot be discounted on the basis of evidence, AT BEST, the study *Meets Evidence Standards with Reservations*. The evidence must demonstrate that the severity of overall attrition is not likely to threaten the internal validity of the study findings.
    - In the case of extremely severe attrition problems, the review Principal Investigator and the review team, in consultation with the WWC Technical Review Team, may make the decision that the study does not meet evidence screens.

- **Differential Attrition**: Differential attrition refers to the situation in which the percentage of the original study sample retained in the follow-up data collection is substantially different for the intervention and the control/comparison groups. Severe differential attrition makes the results of a study suspect because it may compromise the comparability of the study groups. For an RCT to *Meet Evidence Standards,* there should not have been a severe differential attrition problem.

---

4 See the WWC Study Design Classification technical working paper for a description of acceptable randomization problems versus problematic attempts at randomization that would downgrade a study.
5 An RCT trial that is relegated to "Meets Evidence Standards with Reservations" is still eligible for review. For example, if an RCT has severe unaddressed attrition, it would be identified as "Meets Evidence Standards with Reservations." A QED with severe unaddressed attrition, however, would be removed from review.

- If there was severe differential attrition that cannot be discounted on the basis of evidence, AT BEST, the study *Meets Evidence Standards with Reservations*. The evidence must demonstrate that the severity of differential attrition is not likely to threaten the internal validity of the study findings
- In the case of extremely severe attrition problems, the PI and the review team, in consultation with the TRT, may make the decision that the study does not meet evidence screens.

- **Intervention Contamination**: Intervention contamination occurs when something happens after the beginning of the intervention and affects the outcome for the intervention or the control/comparison group, but not both. For an RCT to *Meet Evidence Standards*, there should be no evidence of a changed expectancy/novelty/disruption, a local history event, or any other intervention contaminants.[6]
  - If there is evidence of intervention contamination, the study *Meets Evidence Standards with Reservations*.

- **Mismatch Between Unit of Assignment and Unit of Analysis**: Some RCTs may be designed and implemented well but the analysis of data may be incorrect. A common problem is that the units of random assignment may not match up with the units of analysis and this feature of the study design is ignored in the analysis. Ignoring this fact leads to inflated estimates of the statistical significance of study findings.

  Mismatch does not affect the rating given to a study; that is, it does not affect the statement about meeting evidence standards because the standards rely solely on the design rather than the analysis of the study. Nevertheless, WWC reports need to recognize the mismatch problem when it occurs.

6 Intervention contamination poses a threat to the validity of the evidence for an intervention's effects in that the observed difference between the intervention and the control groups may not be entirely attributable to the intervention, but may reflect the effect of the contaminant.

*If the study appears to use a **quasi-experimental design with equating**, use the following rules to determine whether the study "Meets Evidence Standards with Reservations" or "Does Not Meet Evidence Screens."*

- **Group Assignment**: Studies in which participants were placed into groups using procedures other than random assignment or a cutoff score on a pre-intervention measure are assumed to *Meet Evidence Standards with Reservations*, unless one or more of the following conditions is violated:

- **Baseline Equivalence**: The groups should have been comparable at baseline, or incomparability should have been addressed in the analysis and impact estimates should reflect adjustments for that incomparability.
    - If there is incomparability that is not accounted for in the analysis, the study *Does Not Meet Evidence Screens*.
    - If the groups appeared to be patently incomparable at baseline,[7] and the incomparability is unlikely to be adequately adjusted through statistical procedures, the study *Does Not Meet Evidence Screens.*

- **Overall Attrition**: For a QED to *Meet Evidence Standards with Reservations,* there should not be a severe overall attrition problem or, if there was, it should have been accounted for in the analysis.
    - If there was severe overall attrition that cannot be discounted on the basis of evidence, the study *Does Not Meet Evidence Screens*. The evidence must demonstrate that the severity of overall attrition is not likely to threaten the internal validity of the study findings.

- **Differential Attrition**: For a QED to *Meet Evidence Standards with Reservations,* there should not have been a severe differential attrition problem or, if there was, it should have been accounted for in the analysis.
    - If there was severe differential attrition that cannot be discounted on the basis of evidence, the study *Does Not Meet Evidence Screens*. The evidence must demonstrate that the severity of differential attrition is not likely to threaten the internal validity of the study findings.

- **Intervention Contamination**: There should be no evidence of a changed expectancy/novelty/disruption, a local history event, or any other intervention contaminants.
    - If there is evidence of an intervention contaminant, the study *Does Not Meet Evidence Screens*.

- **Mismatch Between Unit of Assignment and Unit of Analysis**: Some QEDs may be designed and implemented well but the analysis of data may be incorrect. A common problem is that the units of random assignment may not

7 The Principal Investigator and the Review Team for a given topic area have the discretion to determine whether the baseline incomparability in a study was too substantial to be adequately adjusted. The decision rules for handling such studies will be documented and justified.

match up with the units of analysis and this feature of the study design is ignored in the analysis. Ignoring this fact leads to inflated estimates of the statistical significance of study findings.

Mismatch does not affect the rating given to a study; that is, it does not affect the statement about meeting evidence standards because the standards rely solely on the design rather than the analysis of the study. Nevertheless, WWC reports need to recognize the mismatch problem when it occurs.

STAGE 3: IDENTIFYING OTHER IMPORTANT CHARACTERISTICS OF STUDIES THAT MEET EVIDENCE STANDARDS (WITH OR WITHOUT RESERVATIONS)

OVERVIEW

All studies that pass the evidence standards and are rated as either *Meets Evidence Standards* or *Meets Evidence Standards with Reservations* during the second review stage are further reviewed to describe other important study characteristics. The purpose of the Stage 3 review is to collect contextual information about the studies that provide evidence for the effectiveness of the interventions being tested, and to aid the interpretation of the findings presented in the WWC intervention and topic reports. The additional information collected during the third review stage does not affect the ratings of the studies on the evidence standards (i.e., *Meets Evidence Standards*, *Meets Evidence Standards with Reservations*, or *Does Not Meet Evidence Screens)*, which are determined during the second review stage.

OTHER STUDY CHARACTERISTICS

- **Variations in People, Settings, and Outcomes**[8]
  - Subgroup Variation: What subgroups were included in the study?
  - Setting Variation: In what settings did the study take place?
  - Outcome Variation: What outcomes were measured in the study?

- **Analysis of Intervention's Effects on Different Subgroups, Settings, and Outcomes**
  - Analysis by Subgroups: For what subgroups were effects estimated?
  - Analysis by Setting: For what settings were effects estimated?
  - Analysis by Outcome Measures: For what outcomes were effects estimated?

- **Statistical Reporting**
  - Complete Reporting: Are findings reported for most of the important measured outcomes?[9]
  - Formula for Effect Size Computation: Can effect sizes be estimated using the standard formula (or an algebraic equivalent)?

8 Information about the variations in people, settings, and outcomes of the studies as well as information about analysis within subgroups will help to assess the generalizability of the study findings.

9 The purpose of this question is to assess the extent to which the study findings are biased by potential selective reporting, as reported findings are more likely to demonstrate positive intervention effects than findings from the same study that are not reported by the study authors.